



Superstore公司

业务市场可视化分析和预测

数据案例

作者：文星迪、钟豪

目录

CONTENTS

01 案例内容设计

02 案例教学设计

Part 01

第一部分

案例内容设计

重点介绍案例的整体情况

数据集

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer I	Segment	Country	City	State	Postal Cod	Region	Product ID	Category	Sub-Categ	Product Na	Sales	Quantity	Discount	Profit
2	1	CA-2016-	2016-11-0	2016-11-1	Second Cla	CG-12520	Claire Gute	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-BO-1	Furniture	Bookcases	Bush Some	261.96	2	0	41.9136
3	2	CA-2016-	2016-11-0	2016-11-1	Second Cla	CG-12520	Claire Gute	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-CH-1	Furniture	Chairs	Hon Delux	731.94	3	0	219.582
4	3	CA-2016-	2016-06-1	2016-06-1	Second Cla	DV-13045	Darrin Van	Corporate	United Sta	Los Angele	California	90036	West	OFF-LA-1	Office Supp	Labels	Self-Adhes	14.62	2	0	6.8714
5	4	US-2015-	2015-10-1	2015-10-1	Standard C	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South	FUR-TA-1	Furniture	Tables	Bretford Cf	957.5775	5	0.45	-383.031
6	5	US-2015-	2015-10-1	2015-10-1	Standard C	SO-20335	Sean O'Do	Consumer	United Sta	Fort Laude	Florida	33311	South	OFF-ST-1	Office Supp	Storage	Eldon Fold	22.368	2	0.2	2.5164
7	6	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	FUR-FU-1	Furniture	Furnishings	Eldon Expr	48.86	7	0	14.1694
8	7	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-AR-1	Office Supp	Art	Newell 322	7.28	4	0	1.9656
9	8	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	TEC-PH-1	Technolog	Phones	Mitel 5320	907.152	6	0.2	90.7152
10	9	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-BI-1	Office Supp	Binders	DXL Angle	18.504	3	0.2	5.7825
11	10	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	OFF-AP-1	Office Supp	Appliances	Belkin F5C2	114.9	5	0	34.47
12	11	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	FUR-TA-1	Furniture	Tables	Chromcraft	1706.184	9	0.2	85.3092
13	12	CA-2014-	2014-06-0	2014-06-1	Standard C	BH-11710	Brosina Ho	Consumer	United Sta	Los Angele	California	90032	West	TEC-PH-1	Technolog	Phones	Konftel 250	911.424	4	0.2	68.3568
14	13	CA-2017-	2017-04-1	2017-04-2	Standard C	AA-10480	Andrew All	Consumer	United Sta	Concord	North Caro	28027	South	OFF-PA-1	Office Supp	Paper	Xerox 1967	15.552	3	0.2	5.4432
15	14	CA-2016-	2016-12-0	2016-12-1	Standard C	IM-15070	Irene Mad	Consumer	United Sta	Seattle	Washingto	98103	West	OFF-BI-1	Office Supp	Binders	Fellowes Pl	407.976	3	0.2	132.5922
16	15	US-2015-	2015-11-2	2015-11-2	Standard C	HP-14815	Harold Pav	Home Offic	United Sta	Fort Worth	Texas	76106	Central	OFF-AP-1	Office Supp	Appliances	Holmes Re	68.81	5	0.8	-123.858
17	16	US-2015-	2015-11-2	2015-11-2	Standard C	HP-14815	Harold Pav	Home Offic	United Sta	Fort Worth	Texas	76106	Central	OFF-BI-1	Office Supp	Binders	Storex Dur	2.544	3	0.8	-3.816
18	17	CA-2014-	2014-11-1	2014-11-1	Standard C	PK-19075	Pete Kriz	Consumer	United Sta	Madison	Wisconsin	53711	Central	OFF-ST-1	Office Supp	Storage	Stur-D-Stc	665.88	6	0	13.3176
19	18	CA-2014-	2014-05-1	2014-05-1	Second Cla	AG-10270	Alejandro C	Consumer	United Sta	West Jor	Utah	84084	West	OFF-ST-1	Office Supp	Storage	Fellowes St	55.5	2	0	9.99
20	19	CA-2014-	2014-08-2	2014-09-0	Second Cla	ZD-21925	Zuschuss D	Consumer	United Sta	San Francis	California	94109	West	OFF-AR-1	Office Supp	Art	Newell 341	8.56	2	0	2.4824
21	20	CA-2014-	2014-08-2	2014-09-0	Second Cla	ZD-21925	Zuschuss D	Consumer	United Sta	San Francis	California	94109	West	TEC-PH-1	Technolog	Phones	Cisco SPA	213.48	3	0.2	16.011
22	21	CA-2014-	2014-08-2	2014-09-0	Second Cla	ZD-21925	Zuschuss D	Consumer	United Sta	San Francis	California	94109	West	OFF-BI-1	Office Supp	Binders	Wilson Jon	22.72	4	0.2	7.384
23	22	CA-2016-	2016-12-0	2016-12-1	Standard C	KB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AR-1	Office Supp	Art	Newell 318	19.46	7	0	5.0596
24	23	CA-2016-	2016-12-0	2016-12-1	Standard C	KB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AP-1	Office Supp	Appliances	Acco Six-C	60.34	7	0	15.6884
25	24	US-2017-	2017-07-1	2017-07-1	Second Cla	SF-20065	Sandra Flai	Consumer	United Sta	Philadelphi	Pennsylvan	19140	East	FUR-CH-1	Furniture	Chairs	Global Del	71.372	2	0.3	-1.0196
26	25	CA-2015-	2015-09-2	2015-09-3	Standard C	EB-13870	Emily Burn	Consumer	United Sta	Orem	Utah	84057	West	FUR-TA-1	Furniture	Tables	Bretford Cf	1044.63	3	0	240.2649
27	26	CA-2016-	2016-01-1	2016-01-2	Second Cla	EH-13845	Eric Hoffm	Consumer	United Sta	Los Angele	California	90049	West	OFF-BI-1	Office Supp	Binders	Wilson Jon	11.648	2	0.2	4.2224

9994行×21列

02 数据摘要



```
# 查看数据集摘要:  
round(df.describe())
```

	Row ID	Postal Code	Sales	Quantity	Discount	Profit
count	9994.0	9994.0	9994.0	9994.0	9994.0	9994.0
mean	4998.0	55190.0	230.0	4.0	0.0	29.0
std	2885.0	32064.0	623.0	2.0	0.0	234.0
min	1.0	1040.0	0.0	1.0	0.0	-6600.0
25%	2499.0	23223.0	17.0	2.0	0.0	2.0
50%	4998.0	56430.0	54.0	3.0	0.0	9.0
75%	7496.0	90008.0	210.0	5.0	0.0	29.0
max	9994.0	99301.0	22638.0	14.0	1.0	8400.0



基本分析

```
# 计算总销售额:  
print("The total sales of Superstore is ${0}.".format(round(df["Sales"].sum(),2)))  
  
The total sales of Superstore is $2297200.86.
```

```
# 计算总利润:  
print("The total profit of Superstore is ${0}.".format(round(df["Profit"].sum(),2)))  
  
The total profit of Superstore is $286397.02.
```



01 基于类别的分析

营销策划

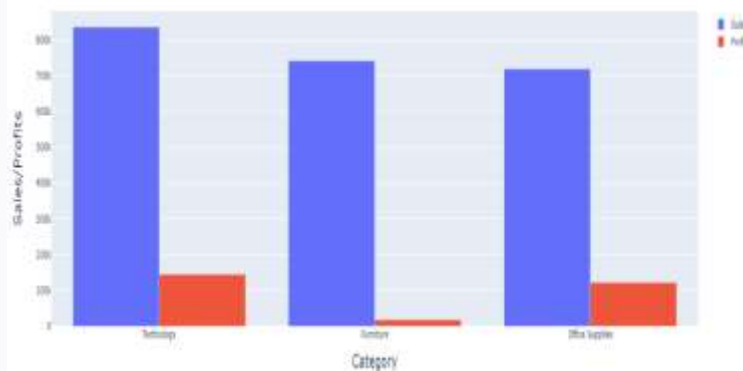
42%



通过箱线图对各个类别商品销售数据的集中趋势进行分析，发现科技大类的产品离散程度最为显著，同时利润接近0的频率也越少，而家具大类离散程度最为集中，但是利润趋近于0的频率极多

用户调研

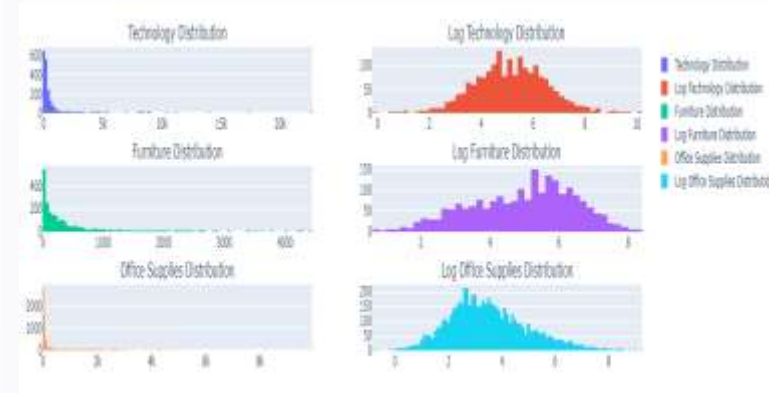
58%



通过直方图得出总结科技大类产品贡献了最多的销售额以及毛利润，但是销售额贡献第二的家具大类却贡献了最低的利润

需求管理

74%



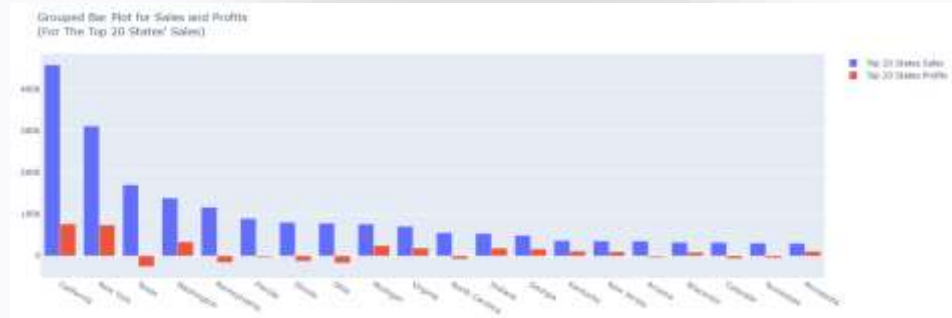
通过对数正态分布来验证三大品类是否为正态分布以及展示其相应分布趋势

基于州的分析

州数据描述

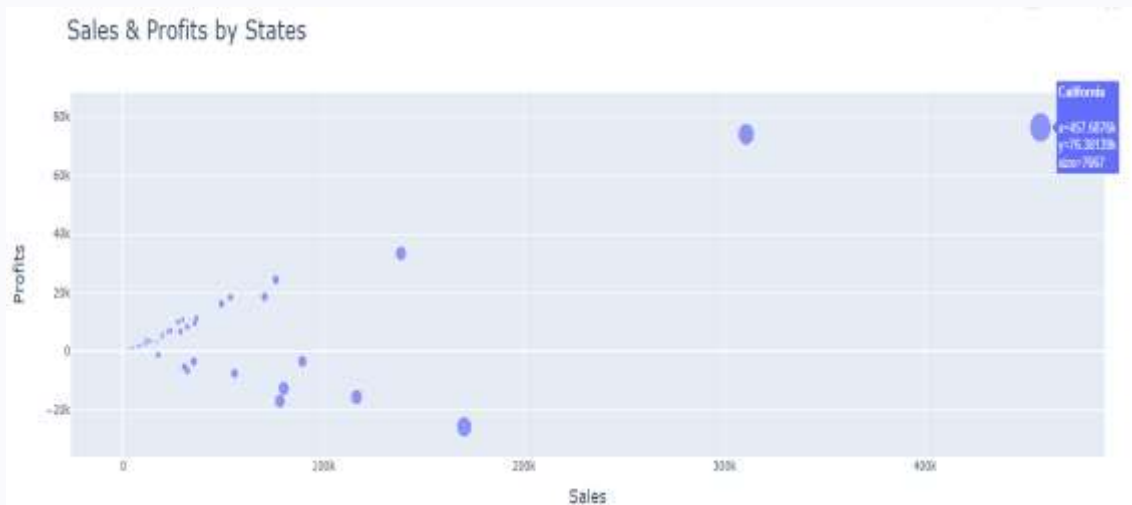
	Postal Code	Sales	Quantity	Discount	Profit
State					
Alabama	2195669	19510.640000	256	0.000000	5786.825300
Arizona	19102126	35282.001000	862	68.000000	-3427.924600
Arkansas	4339309	11678.130000	240	0.000000	4008.687100
California	184382639	457687.631500	7667	145.600000	76381.387100
Colorado	14613828	32108.118000	693	57.600000	-6527.857900
Connecticut	531005	13384.357000	281	0.600000	3511.491800
Delaware	1896504	27451.069000	367	0.600000	9977.374800
District of Columbia	200160	2865.020000	40	0.000000	1059.589300
Florida	12640225	89473.708000	1379	114.650000	-3399.301700
Georgia	5685480	49095.840000	705	0.000000	16250.043300
Idaho	1752709	4382.486000	64	1.800000	826.723100
Illinois	29873772	80166.101000	1845	191.900000	-12607.887000
Indiana	6991602	53555.360000	578	0.000000	18382.936300
Iowa	1537707	4579.760000	112	0.000000	1183.811900
Kansas	1603798	2914.310000	74	0.000000	836.443500
Kentucky	5725336	36591.750000	523	0.000000	11199.696600
Louisiana	2972649	9217.030000	156	0.000000	2196.102300
Maine	34725	1270.530000	35	0.000000	454.486200
Maryland	2206740	23705.523000	420	0.600000	7031.178800
Massachusetts	268295	28634.434000	491	2.100000	6785.501600

州数据分析



- 热力图高效直观地反映出销售额利润最高、最低的州分别是California 和 Lllinois
- 最为通用的条形图清晰地展示出Top 20 利润为负的大州以及最高销售额的大州

基于州的分析

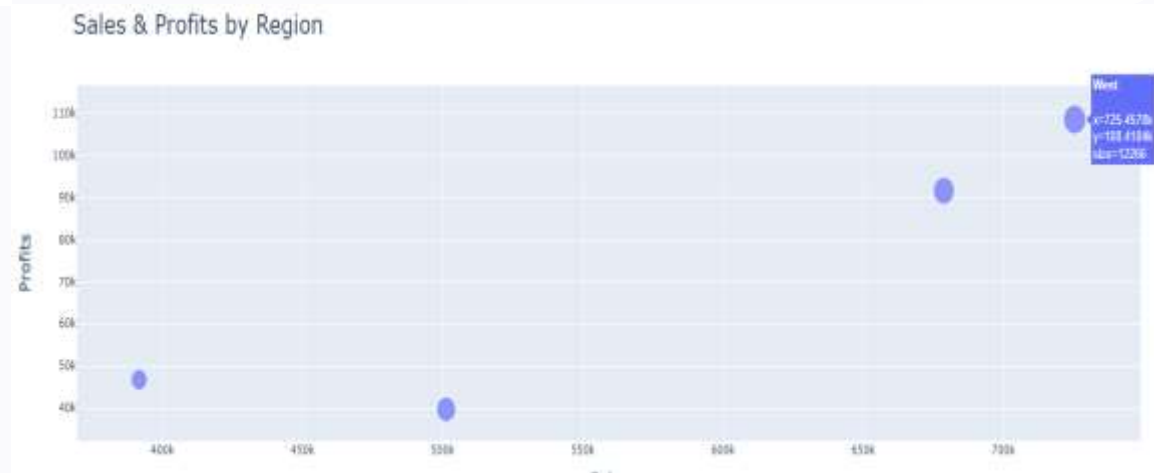


大州分析

散点图分析利润贡献最高的Top 3大州分别为 Californian, New York and Washington, 利润额亏损最多的州是Texas

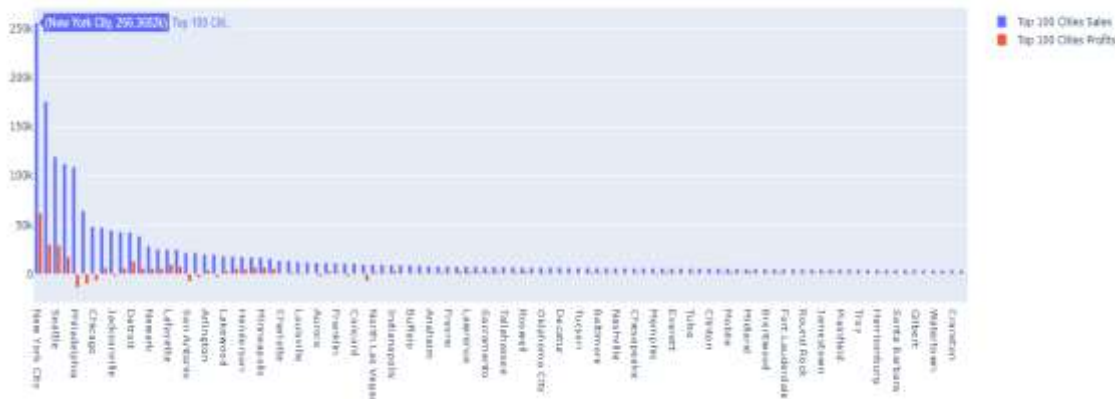
方位分析

利润额贡献最高的分别为西部其实是东部，最低的为中部地区



基于城市的分析

Grouped Bar Plot for Sales and Profits
(For The Top 100 Cities' Sales)



城市分析

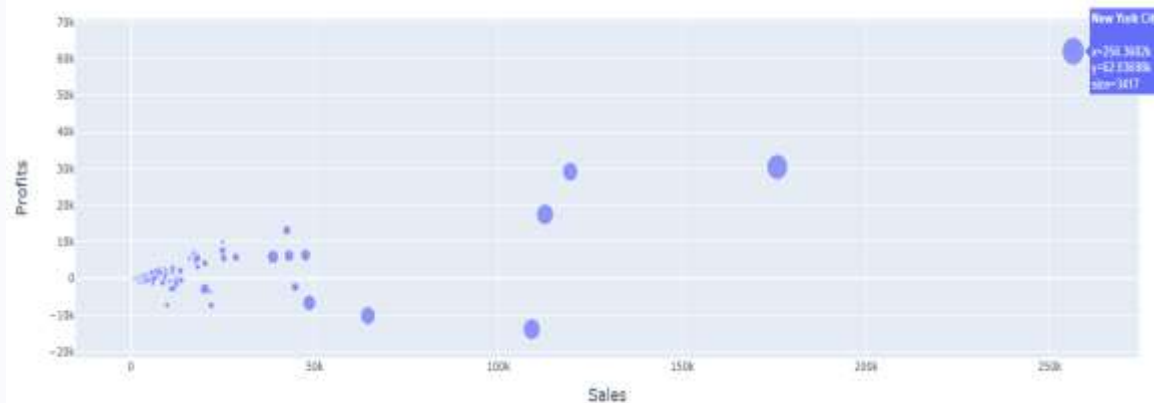
通过柱状图可以发现销售额和利润额最高的城市为New York City，其次是洛杉矶，同时还可以观察到一些城市出现利润为负的现象



图表多元化

通过气泡图同时印证了柱状图所表示的现象，进行多元化的交叉验证,展现不同图表相同场景的差异性、适用性

Sales & Profits by Cities

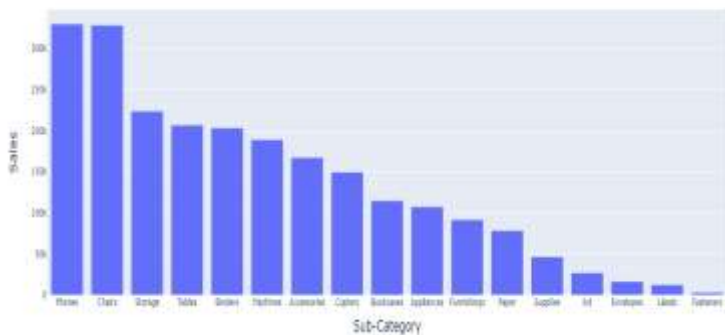


01 基于子类别的分析

42%

初步比较

Bar Chart for the Sales of each Sub-Category

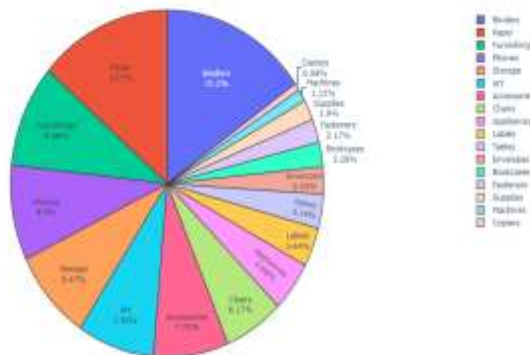


通过条形图对贡献销售额的产品进行高到低的排序，手机和椅子并列第一，安全带贡献最少

58%

整体观察

Pie Chart for Total Sales of each Sub-Category



通过饼图进行整体观察确定活页夹、纸张、穿戴用品、手机等位列前茅，多为消费品撑起销售额

74%

分析综述

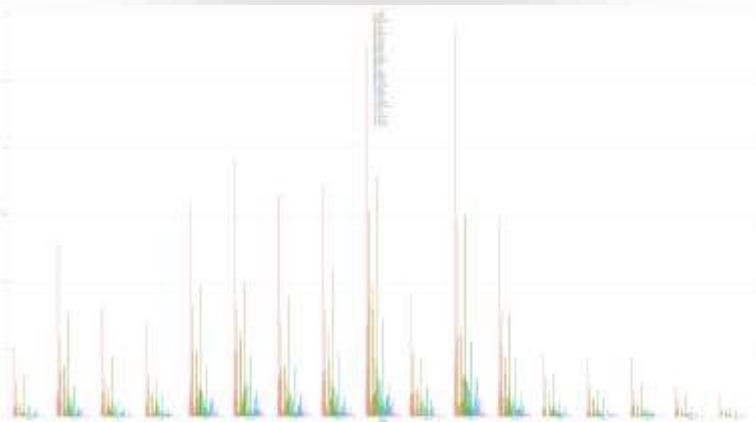


通过条形统计图进行初步比较子类别的销售额，科技小类总体利润、销售额贡献最高，但是销售数量最少

01 基于子类别的分析

42%

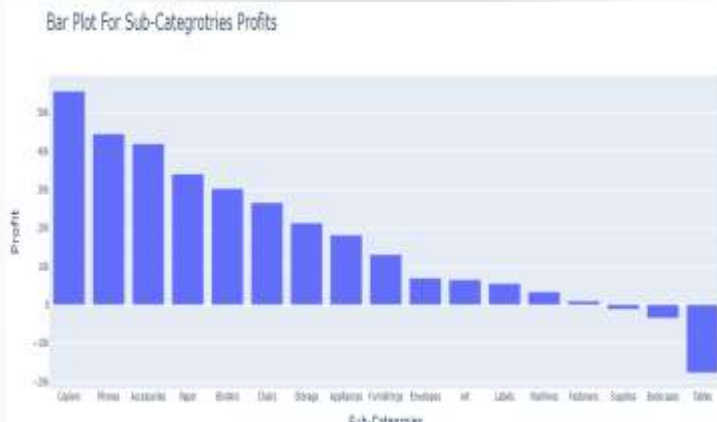
初步比较



从条形统计图中我们可以看到钩扣、补给品、机器和复印机等产品在加利福尼亚州大卖特卖

58%

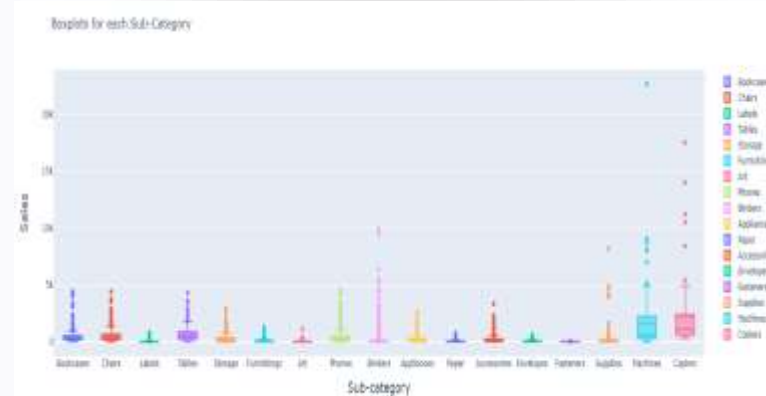
整体观察



通过条形图进行利润观察有桌子贡献最多的亏损，复印机、手机等科技类产品贡献最多的利润

74%

分析综述



通过箱线统计图对子类别的集中趋势如中位数、四分位线、四分位距等进行分析综述

TensorFlow 神经网络模型 预测

📄 模型训练

- 使用 TensorFlow 2.0 框架，搭建了一个简单的全连接神经网络模型，解决了 收入与其他变量之间的关系
- 模型编译中，优化器选择为 adam，损失函数为 mse（均方误差）。
- 模型训练100次



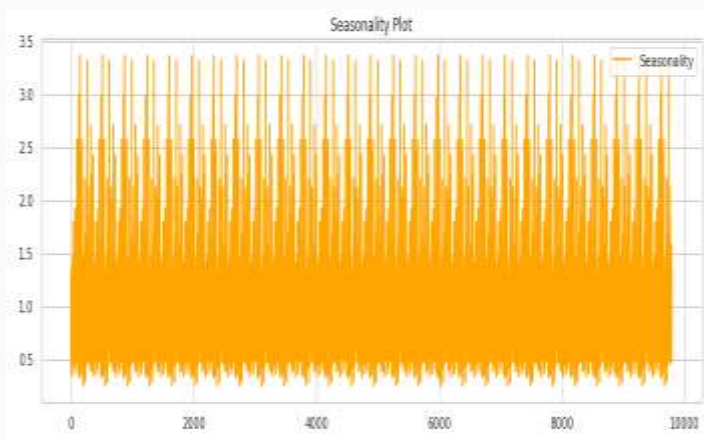
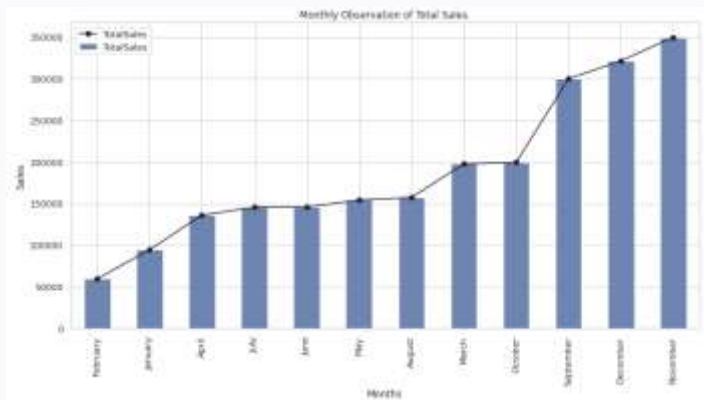
100次

```
Epoch 1/100  
172/172 [=====] - 7s 40ms/step - loss: 454819.4375 - val_loss: 173004.8906 - lr: 0.0010  
Epoch 2/100  
172/172 [=====] - 7s 39ms/step - loss: 349885.1562 - val_loss: 186095.0938 - lr: 0.0010  
Epoch 3/100  
172/172 [=====] - 7s 39ms/step - loss: 234907.0625 - val_loss: 1582909.0000 - lr: 0.0010  
Epoch 4/100  
172/172 [=====] - 7s 39ms/step - loss: 161334.9062 - val_loss: 5310099.0000 - lr: 0.0010  
Epoch 5/100  
172/172 [=====] - 7s 39ms/step - loss: 101324.8281 - val_loss: 10487272.0000 - lr: 0.0010  
Epoch 6/100  
172/172 [=====] - 7s 40ms/step - loss: 62704.3320 - val_loss: 13444157.0000 - lr: 0.0010
```

Test Loss: 216452.57812

Test R² Score: 0.25251

季节性分析模型预测

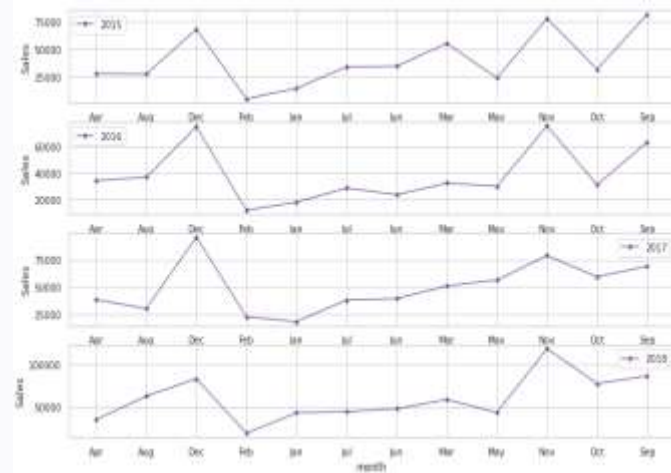


月度趋势

通过观察月度总销售额发现9月到12月之间增幅骤升

用户数据

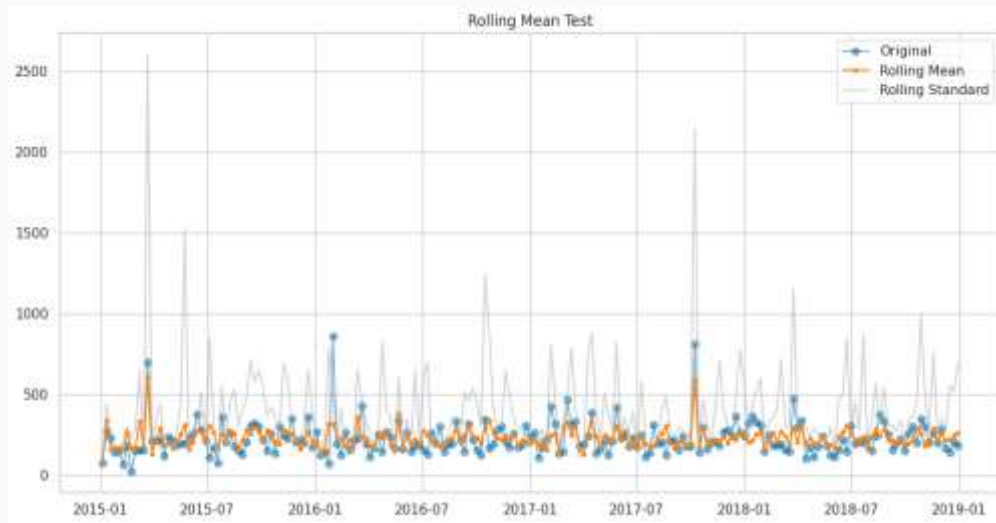
通过同轴折线图对历年月度变化进行观察，可以发现每年9月、11月、12月销售额都出现不同幅度的上升



时间序列模型 预测

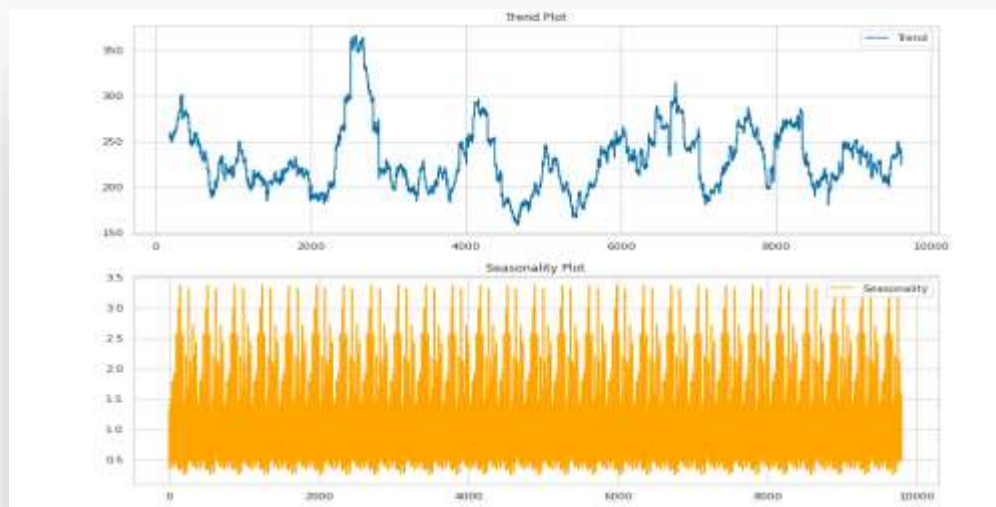
滚动平均测试

- 右图表明，平均值和标准差随时间变化不大，这意味着平均值和方差是恒定的，**时间序列平稳**，符合进行**ARIMA模型预测**的条件



产品业绩

- 右图，没有显示任何数据有关时间的趋势，未来数据**不会受时间因素的影响**，它的行为并不会随着时间的推移而变化，因此**不需要进行对时间数据矩阵的差分**，继续施行**差分整合移动平均自回归预测模型**



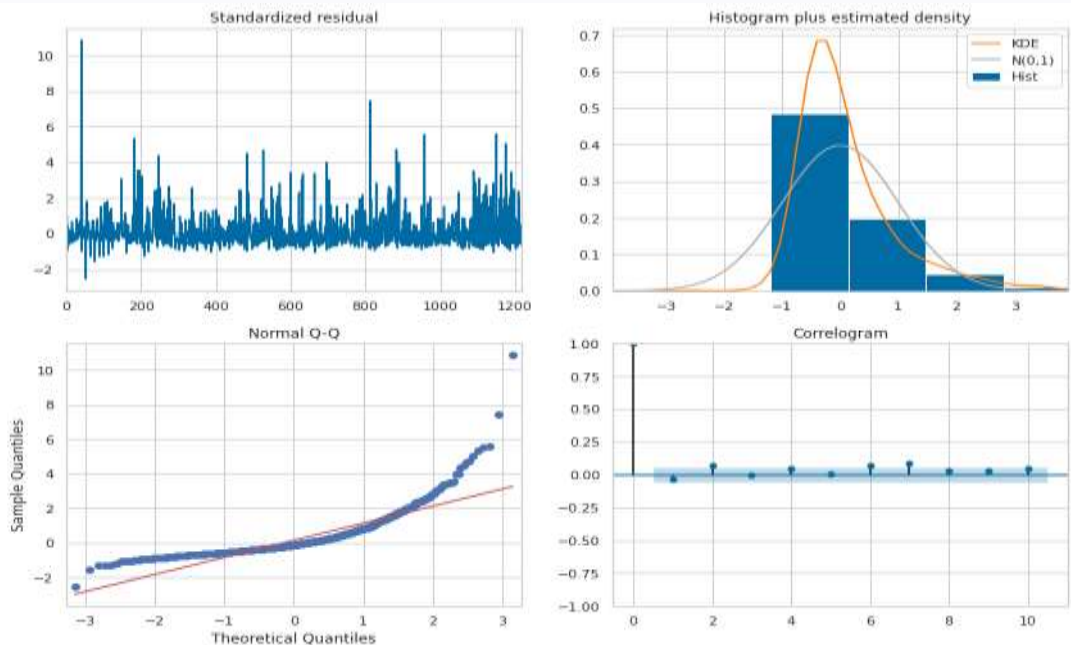
差分整合移动平均 自回归预测模型

模型初判

- 通过图一折线图可以看出残差大表明观察值和预测值之间的差值大，波动大离散程度高表明预测准确的稳定性低
- 图二柱状曲线图的红色KDE线和灰色N(0,1)线偏差大，表明残差是不服从正态分布的。

模型检验

- 通过图三残差散点图和趋势线，发现残差序列非正态分布，不存在高斯白噪声，通过了白噪声检验，没有信息可以继续提取，建模终止。
- 图四来判断自相关和偏相关的拖尾性或者截尾性



42%

产品优化

71万

数据转化

58万

经济效益

XGboost 回归 预测模型



XGboost 回归预测优势

XGBoost的基本思想和GBDT相同，可以说是极优化的GBDT，使二阶导数使损失函数更精准；正则项避免树过拟合；Block存储可以并行计算等。

1. GBDT

GBDT模型

GBDT是一种基于boosting集成思想的加法模型。

GBDT训练

采用前向分布算法进行贪婪的学习，每次迭代都学习一棵CART树，来拟合之前($t-1$)棵树的预测结果与训练样本真实值的残差。

> 用负梯度近似模拟残差

GBDT缺点

依赖强，并行难，效率低。



RMSE

SARIMAX 2404.877287

XGBRegressor 1714.644163

XGboost回归预测模型的均方根误差小于差分整合移动平均自回归模型。此案例中使用 XGboost回归预测模型来预测销售额更准确。

Part 02

第二部分

案例教学设计

该数据案例如何应用于教学

案例教学目的

非计算机相关专业同学着重了解对业务领域的分析

- 了解数据科学背景下编程的主要原则和工具
- 培养基本处理和可视化数据的能力
- 学习数据可视化的基本图形和原则
- 引导将分析性思维应用于商业领域
- 引导将案例应用于不同的分析场景




计算机相关专业同学了解更多编程技术的知识


- 学习Python编程基本语法
- 学习Python中数据结构的类型
- 学习Python主要绘图包的使用
- 学习Python中的变量、可变性和别名
- 学习Python中的异常和错误处理
- 学习编程中的调试
- 学习数据整理操作
- 在Python中与机器学习框架进行交互


案例教学要素


运行环境


 7, 8 或 10 64位系统  OS X 10.13 或更高版本 ✓

 Intel Sandy Bridge, SSE 4.2, or 更新 CPUs ✓

 上行、下行速率至少50Mbps的宽带 ✓

 至少500M的空闲硬盘空间 ✓

 8GB以上内存 ✓

 系统完全管理员权限 ✓

实验内容

- 对现有公司的销售数据集进行探索性分析和预测性分析
- 在运行中强化学生学习到的基本编程知识、描述统计知识、机器学习知识

案例教学步骤

分8个模块进行讲解，前7个模块每个模块使用1~2节课讲解知识，提供1节实操练习课让学生对所学技能进行练习。第8模块演示数据案例。

	模块	教授内容
数据科学生态系统和编程工具	1	<ul style="list-style-type: none">• 介绍数据科学并回顾真实世界的的数据示例。• 介绍使用基本编程工具和技术经验，例如Jupyter Notebook、IDE和Git。
与数据结构进行交互	2	<ul style="list-style-type: none">• 熟悉各种数据类型和结构以及流行的数据交换格式（例如JSON、XML、CSV）。• 使用Python处理各种数据类型和结构以及数据交换格式。

案例教学步骤

	模块	教授内容
核心编程概念	3	<ul style="list-style-type: none">理解并使用Python中的基本编程概念，如循环、变量和函数。理解并使用Python中的基本编程概念，如异常、错误处理、测试和调试。
数据整理	4	<ul style="list-style-type: none">理解并使用Python中的数据类型和结构。

案例教学步骤

	模块	教授内容
图形和数据可视化	5	<ul style="list-style-type: none">• 熟悉Python中的图形和数据可视化。• 理解图形范例的语法及其在Python中的实现。• 生成网络可视化。
机器学习框架	6	<ul style="list-style-type: none">• 介绍机器学习框架。• 获得形成数据分析管道和并行计算原理的经验。• 与Python中的机器学习框架交互。
软件开发	7	<ul style="list-style-type: none">• 获得记录代码的经验。• 了解软件测试框架和测试驱动开发。• 开发Python包。
数据案例讲解	8	<ul style="list-style-type: none">• 进行数据案例的运行演示。

案例教学步骤

考核方式

每个模块布置1项实操作业，学生需将作业提交到指定平台上，老师进行打分。在教学过程结束后安排一场闭卷考试检验学生学习情况。

平常作业总共占50%的最终成绩，考试占50%的最终成绩。



案例教学要求

老师教授

- 确保非计算机相关专业同学能够独立替换案例中的变量和数据集
- 确保计算机相关专业同学能够独立写出达到分析目的的代码

学生学习

- 非计算机相关专业的同学能够使用案例中提供的技术分析不同的数据集
- 计算机专业同学深入理解案例中涉及到的编程原理

案例试验结果

结论分析

Superstore公司的总销售额和利润都很高。利润约占总销售额的12%。部分州和城市出现了亏损，但总体而言仍在盈利。在所有类别中，科技类商品销售最多，利润最多。在所有子类别中，订书机占有最多的销售额。在所有的州中，加利福尼亚州是销量最多的州，那里是Superstore公司的主要收入市场。纽约市是销量最多的城市。此外，住在美国西部的人比东部的人更倾向于从Superstore订购更多的商品。

从各大类别和子类别的销售数据及其位置分布的分析中发现：

- 1.该公司可以尝试扩大美国中部州的销售份额占比，由于公司在东部各州（如纽约州）和西部的加州销售已经达到饱和，而中部各州销售份额占比仅占15.2%，具有较大盈利空间。
- 2.该公司的三大类商品：科技类产品、家具产品以及办公室用品中，科技产品和办公室用品利润率相对于家居产品较高，盈利能力较强。因此该公司可以加大对新型科技产品和办公室用品的研发投入比重，让资源用到更能带来价值的地方。
- 3.在该公司销售产品的子类别中，订书机、办公室用纸、手机、室内陈设以及库存用品占据了总销售额的55.85%，公司在保持原有子产品竞争力的同时，可以加大一些销售额占比不高但利润率高的科技产品（如衣物配饰、机器和打印机）的市场投入，这些产品具有较大潜在价值，能创造更多的利润。

通过对每年销售额趋势分析，可以看出每年第3~4季度的销售额都具有增长趋势，可以预测公司未来销售额具有逐年增长潜力。

案例教学方法

课堂教学理论+实操结合

课程作业

- 个案研究和数据分析
- 10页报告(无代码)
- RMarkdown和Jupyter笔记本中的Python代码

期末考试

- 2小时闭卷书面考试
- 考察理论知识、代码应用



先进教学手段



在线运行环境



Azure Notebooks
Jupyter For The Cloud

kaggle



本地运行环境



Visual Studio Code



可扩展性

读取数据集

```
df = pd.read_csv("Superstore (16~19).csv")
```

季节性销售趋势分析

```
1. # 由于数据集是时间序列数据，因此先将其转换为时间序列数据。
data['order_date'] = pd.to_datetime(data['order_date'], dayfirst=True)
data['ship_date'] = pd.to_datetime(data['ship_date'], dayfirst=True)

2. # 提取"年月"、"季"和"月"列
data['yearmonth'] = data['order_date'].apply(lambda x: x.strftime("%Y-%m"))
data['year'] = data['order_date'].dt.year
data['month'] = data['order_date'].dt.month

3. # 提取记录销售产品所属的类别序列
data['ship_date'] = data['ship_date'].dt.date
data['order_date'] = data['order_date'].dt.date

# 查看前5行数据
data.head()
```



每个城市销售额和利润的散点图：

```
fig = go.Figure(data=px.scatter(x=temp_data.groupby('City').sum()['Sales'],
                                y=temp_data.groupby('City').sum()['Profit'],
                                hover_name=temp_data.groupby('City').sum().index,
                                size=temp_data.groupby('City').sum()['Quantity']))

fig['layout']=go.Layout(title='Sales & Profits by Cities',
                        titlefont=dict(size=25),
                        xaxis=dict(title='Sales', titlefont=dict(size=18)),
                        yaxis=dict(title='Profits', titlefont=dict(size=18)))

iplot(dict(data=fig))
```

感谢观看